RA-GCN：抑制过平滑现象的文本分类算法 *

苏凡军，马明旭†，佟国香

(上海理工大学 光电信息与计算机工程学院，上海 200093)

摘 要：现有大多数利用图神经网络的算法进行文本分类时，忽略了图神经网络的过度平滑问题和由于文本图拓扑差异引入的误差，导致文本分类的性能不佳。针对这一问题，提出了衡量多个文本图表示的平滑度的方法 WACD 以及抑制过平滑现象的正则项 RWACD。随后提出了基于注意力和残差的网络结构 ARS，用于弥补由于图拓扑差异引起的文本信息的损失。最后，提出了图卷积神经网络文本分类算法 RA-GCN。RA-GCN 在图表示学习层使用 ARS 融合文本表示，在读出层使用 RWACD 抑制过平滑现象。在 6 个中英文数据集上进行实验，实验结果证明了 RA-GCN 的分类性能，并通过多个对比实验验证了 RWACD 和 ARS 的作用。

关键词：文本分类；图卷积神经网络；过平滑；注意力机制

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2022.02.0037

RA-GCN: text classification algorithm suppressing over-smoothing phenomenon

Su Fanjun, Ma Mingxu†, Tong Guoxiang

(*School of Optical-Electrical & Computer Engineering*, *University of Shanghai for Science & Technology Shanghai* 200093, *China*)

Abstract: Most existing text classification algorithms based on graph neural network ignore the problem of over-smoothing, and ignore the problem of information loss due to graph topology, resulting in poor classification performance. To solve this problem, this paper proposed a method to measure the smoothness of multiple text graph representations WACD and a regularization term RWACD to suppress over-smoothing. Subsequently, this paper proposed an attention and residual-based network structure ARS to compensate for the loss of textual information due to graph topology differences. Finally, this paper proposed a graph convolutional neural network text classification algorithm RA-GCN. RA-GCN used ARS to fuse text representations in the graph representation learning layer, and used RWACD in the readout layer to suppress over-smoothing. This paper conducted experiments on 6 Chinese and English datasets. The experimental results demonstrate the classification performance of RA-GCN, and the effects of RWACD and ARS are verified through multiple comparative experiments.

Key words: text classification; graph convolutional network; over-smoothing; attention mechanism

0 引言

文本分类作为自然语言处理领域的基础问题，已被应用于许多现实场景，例如垃圾邮件检测，新闻分类，情感识别等。文本分类模型的性能很大程度上取决于文本表示的质量。基于深度学习的方法避免了人工设计规则和特征，自动学习语义上有意义的表示[1]。基于 CNN 和 RNN 的深度学习方法可以很好地捕获局部连续序列中的语义和句法特征，但对非连续词和长距离语义信息的提取仍然存在限制[2~4]。

近年来，图神经网络缓解了上述现象。Yao[5]构建整个语料库的单张文本-单词异构图，使用 GCN[6]学习词共现信息，更新文本、单词表示，进行文本分类。Wu[7]通过去除非线性激活函数和折叠连续层之间的权重矩阵，将 GCN 简化为 SGC，并且在基于单张语料库异构图的数据上取得了不错的文本分类效果。但是，基于单张异构图的方法不利于测试新的文本，消耗了大量的内存空间，为此，Huang[8]构建每个文本的图数据，共享全局单词表示和边的权值，更好的捕捉局部特征和减少内存消耗。Zhang[9]为提高图方法的归纳学习能力，构建每个文本独特的图数据，使用 GGNN[10]更新单词特征，获取文本表示及类别。但是，上述方法忽略了图神经网络的过度平滑问题，本文关注基于每个文本图表示的图分类[8,9,11]方向，缓解过度平滑现象，提升文本分类性能。

在 GNN 发展过程中，Li[12]首次引起对过度平滑的关注[13]，证明了图卷积是一种特殊的拉普拉斯平滑，并且得出结论：对节点进行平滑操作是 GCN 工作的关键机制，但是执行多次拉普拉斯平滑后，节点特征会收敛至相似值，这个现象被称为过度平滑现象，也被称为过平滑，过平滑会导致节点之间无法区分，从而损害网络性能。Chen[14]验证了平滑是 GNN 的本质，给出了衡量平滑度的方法 MAD(Mean Average Distance)，从图拓扑角度分析了过度平滑的原因，认为信息和噪声的过度混合是影响过度平滑的一个关键因素，提出了抑制过平滑的正则项 MADreg 和迭代训练算法 AdaGraph。同时有研究者提出通过模型优化人为构造的图拓扑，提升模型性能，抑制过度平滑现象。Wang[15]通过多跳注意力机制扩大节点的感受野，使不直接连接但相聚多跳的节点之间进行远程交互，过滤高频噪声信息。Yang[16]利用指针网络[17]寻找多阶邻域中的相关节点，使用一维卷积提取高级特征，过滤噪声信息，减轻过度平滑问题。在网络结构方向，文献[18]借助残差、密集连接和扩张卷积堆叠深层 GCN，显著提高了 GCN 在点云语义分割任务中的性能，缓解了过平滑现象。在数据方向，Rong[19]在每个训练期间随机丢弃图中一定比例的边，以充当数据增强器和消息传递减速器，降低过平滑的收敛速度。

收稿日期：2022-02-09；**修回日期：**2022-03-22　**基金项目：**国家重点研发计划项目(2018YFB1700902)

作者简介：苏凡军(1976-)，男，山东泰安人，讲师，博士，主要研究方向为推荐算法、图神经网络、计算机网络；马明旭(1997-)，男(通信作者)，山东泰安人，硕士研究生，主要研究方向为自然语言处理(mingxuma@126.com)；佟国香(1968-)，女，四川成都人，副教授，硕导，博士，主要研究方向为嵌入式系统开发、图像处理、数据挖掘.

根据文献[12,14]及本文的实验现象，可以发现使用 GCN 进行文本图表示学习时，平滑使得单词特征收敛至相似值，单词表示不可避免地变得相似，损害了文本分类的性能。为此，本文针对文本分类问题，为了更好的衡量及分析单词节点的平滑度，提出了衡量多个文本图表示的平滑度的方法加权平均余弦距离 WACD(Weighted Average Cosine Distance)。WACD 与 MAD[14]不同，MAD 适用于单张图，WACD 则作用于多个图，更适用于本文关注的图分类方向。本文借鉴节点分类中抑制过平滑的方法，在 WACD 的基础上提出了抑制过平滑的正则项 RWACD(Regularization based on Weighted Average Cosine Distance)。随后提出了基于注意力和残差的网络结构 ARS(Attention-based Residual Network Structure)，弥补由于图拓扑差异引起的文本信息损失。与[14~16]不同，ARS 无须迭代训练和寻找重要相关节点，仅使用注意力机制和残差结构，加快训练速度。最后，提出了图卷积神经网络文本分类算法 RA-GCN(RWACD-ARS based Graph Convolutional Neural Network Text Classification Algorithm)。算法 RA-GCN 在图表示学习层使用 ARS 融合文本表示，在读出层使用 RWACD 抑制过平滑现象。实验在 6 个中英文数据集上进行，实验结果证明了 RA-GCN 的性能，并通过多个对比实验验证了 RWACD 和 ARS 的作用。总体来说，本文有以下创新点：

a) 提出了衡量多个文本图表示的平滑度的方法 WACD，并提出了抑制过平滑现象的正则项 RWACD。

b) 提出了基于注意力机制和残差的网络结构 ARS，弥补由于图拓扑差异引起的文本信息损失，同时抑制过平滑现象。

c) 提出了基于 RWACD 和 ARS 的图卷积神经网络文本分类算法 RA-GCN，在 6 个中英文数据集上的实验结果证明了 RA-GCN 的性能。

d) 多方面进行对比实验，验证了 RWACD 和 ARS 均能抑制过平滑现象和提升模型性能；证明了从图拓扑角度弥补文本信息损失决策的正确；分析并探讨了在本文关注的基于每个文本图表示的图分类方向中的过平滑现象。

1　相关研究

本文所提算法针对于文本图分类中的过平滑问题，是对文献[14]算法的改进和完善，因此本节重点介绍文献[14]。文献[14]主要针对节点分类领域的过平滑现象，提出了衡量图表示的平滑度的方法 MAD，抑制过平滑的正则项 MADreg 和迭代训练算法 AdaGraph。

1.1　MAD 与 MADreg

MAD 是基于余弦距离衡量图表示的平滑度的方法。给定图表示矩阵 $H \in R^{n \times d}$，其中 n 为节点数，d 为特征维度。通过余弦距离计算距离矩阵 D，每个节点对之间的距离计算为

$$D_{ik} = 1 - \frac{H_{i,:} \cdot H_{k,:}}{|H_{i,:}| \cdot |H_{k,:}|} \qquad i,k \in [1,2,\dots,n] \tag{1}$$

其中 $H_{i,:}$为图表示 H 的第 i 行。使用余弦距离的原因是余弦距离不受节点向量绝对值的影响，从而更好地反映了图表示的平滑性[14]。

为了得到目标节点对之间的余弦距离，构造目标掩码矩阵 M^{tgt}，得到目标节点对的距离矩阵，计算为

$$D^{tgt} = D \circ M^{tgt} \tag{2}$$

其中，$\circ$ 表示逐元素乘法，$M^{tgt} \in \{0,1\}^{n \times n}$，当 (i,k) 是目标节点对时，$M^{tgt}_{ik} = 1$。然后计算每行非零值的平均值：

$$\bar{D}^{tgt}_i = \frac{\sum_{k=1}^{n} D^{tgt}_{ik}}{\sum_{k=1}^{n} 1(D^{tgt}_{ik})} \tag{3}$$

$$1(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \end{cases} \tag{4}$$

通过计算 $\bar{D}^{tgt}_i$ 中非零值的平均值，得到给定目标节点的 MAD^{tgt}，计算为

$$\mathrm{MAD}^{tgt} = \frac{\sum_{i=1}^{n} \bar{D}^{tgt}_i}{\sum_{i=1}^{n} 1(\bar{D}^{tgt}_i)} \tag{5}$$

文献[14]观察到在节点分类中，拓扑距离小的两个节点更有可能属于同一类别，因此提出了利用图拓扑来近似节点类别，并计算远程和邻居节点的 MAD 差值来估计图表示的过平滑度 MADGap，计算为

$$\mathrm{MADGap} = \mathrm{MAD}^{rmt} - \mathrm{MAD}^{neb} \tag{6}$$

其中，MAD^{rmt} 是图拓扑中远程节点的 MAD 值，MAD^{neb} 是邻居节点的 MAD 值。将 MADGap 引入系数 λ 后得到抑制过平滑的正则项 MADreg，计算为

$$\mathrm{MADreg} = -\lambda \times \mathrm{MADGap} \tag{7}$$

1.2　Adagraph

文献[14]观察到在利用真实标签优化图拓扑时，缓解了过平滑现象，提升了节点分类的性能，因此提出了优化图拓扑的迭代训练算法 AdaGraph。首先训练 GNN，然后根据预测结果删除类间边和添加类内边优化图拓扑，多次执行该过程后，降低了图拓扑差异，抑制了过平滑现象，提升了节点分类的性能。

2　本文研究的算法

MAD、MADreg 与 AdaGraph 适用于基于单张图表示学习的工作，然而本文关注的是基于多个文本图表示的图分类方向，因此文献[14]并不能直接用于本文关注的方向，并且 MADreg 需要寻找最优阶数计算 MADGap，AdaGraph 需要迭代训练优化图拓扑，增加了训练时间，与本文关注的方向有较大差异。

为此，提出了衡量多个文本图表示的平滑度的方法加权平均余弦距离 WACD 及抑制过平滑的正则项 RWACD。提出了基于注意力和残差的网络结构 ARS，弥补由于文本图拓扑差异引起的信息损失，同时抑制过平滑现象。最后，提出了图卷积神经网络文本分类算法 RA-GCN。

2.1　WACD 与 RWACD

WACD 衡量多个文本图的平滑度，值越高表示平滑度越低，过平滑概率越低，反之平滑度越高，过平滑概率越大。

首先，对于单个文本图表示 $H_T \in R^{m \times d}$，其中 m 为单词节点数，d 为词嵌入维度。将所有单词对视为目标节点，利用式(1)~式(5)计算文本图的平均余弦距离 ACD(Average Cosine Distance)。利用每个文本的长度计算 ACD 的加权系数 μ_i，以更好的估计多个文本图表示的平滑度 WACD，计算过程为

$$\bar{l} = \frac{1}{b}\sum_{i=1}^{b} l_i \tag{8}$$

$$\mu_i = \frac{l_i}{\bar{l}} \tag{9}$$

$$\mathrm{WACD} = \frac{1}{b}\sum_{i=1}^{b} \mu_i \times \mathrm{ACD}_i \tag{10}$$

其中，b 表示文本数量，l_i 为第 i 个文本的长度。正则项 RWACD 计算为

$$\mathrm{RWACD} = 1 - \mathrm{WACD} \tag{11}$$

基于文本长度加权平均 ACD 得到 WACD，更好的衡量多个文本图的平滑程度；RWACD 通过降低文本图表示的平滑度降低过平滑的概率。与 MADreg[14]相比，RWACD 无需寻找最优阶数，更适用于本文关注的文本图分类方向。

2.2　ARS

参考节点分类领域对图拓扑方向的探讨[14~16]，本文认为人为构造的文本图拓扑与潜在真实文本拓扑存在偏差，造成了图表示学习中的文本信息损失。因此本文提出，对于每个

网络层，利用注意力机制和残差的网络结构 ARS 缓解上述现象，同时抑制过平滑问题。与[14~16]不同的是，ARS 无须迭代训练和寻找重要相关节点，仅使用注意力和残差，加快训练速度，更适合本文关注的方向。ARS 将在 2.3.2 节中详细介绍。

2.3 RA-GCN

如图 1 所示为 RA-GCN 算法的框架图。为了使框架更加清晰，部分框架使用了红、蓝、绿三种颜色突出计算流程，其中红色表示 GCN 的前向计算流程；蓝色表示 ARS 的前向计算流程；绿色表示 RWACD 的前向计算流程。总的来说，RA-GCN 可分为三个部分，分别为文本处理层、图表示学习层和读出层。文本处理层主要对文本进行处理，转换为图表示学习层的输入。图表示学习层学习文本表示，主要由 GCN 和 ARS 两部分构成，GCN 学习图级别的文本表示，ARS 弥补由于文本图拓扑差异引入的信息损失。读出层获取文本类别，使用交叉熵函数计算损失，使用 RWACD 抑制过平滑。下面详细介绍算法的各个部分以及流程。

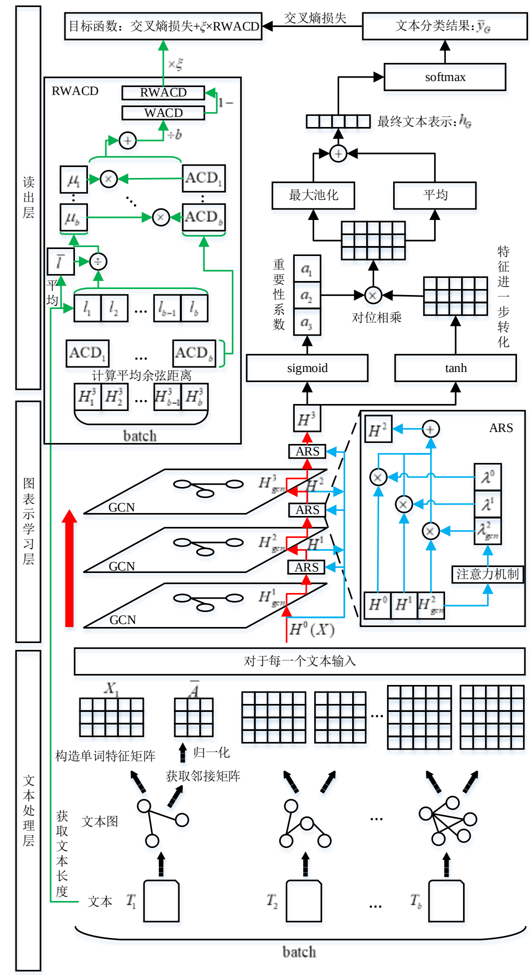


图 1　RA-GCN 算法的框架

Fig. 1　The framework of the RA-GCN algorithm

2.3.1 文本处理层

如图 1 所示，对于文本 $T=\{w_1,w_2,...,w_n\}$，w_i 为单词，文本图数据表示为 $G=(V,E,X)$，$V=\{v_1,v_2,...,v_m\}$ 为唯一出现的单词节点集，$|V|=m$ 为单词个数，$m \le n$，E 为边集，X 为初始单词特征矩阵。使用滑动窗口构建单词节点集 V 和边集 E，边集 E 通过邻接矩阵 A 展现，$A=[a_{ik}]\in R^{m\times m}$，其中 $a_{ik}=1$ 表示单词节点 v_i 和 v_k 相连，0 表示不相连。构建邻接矩阵 A 的度矩阵 $D_T=diag(d_1,d_2,...,d_m)$，其中 d_i 是节点 v_i 的度。归一化的邻接矩阵定义为 $\bar{A}=D_T^{-1/2}AD_T^{-1/2}$。初始单词特征矩阵 $X\in R^{m\times d}$ 使用预训练词嵌入构建，其中 d 是词嵌入维度。

2.3.2 图表示学习层

如图 1 所示，图表示学习层分为 GCN 和 ARS 两个部分，GCN 学习单词共现信息，获取文本图表示；ARS 使用注意力机制和残差结构得到当前图表示学习层的文本表示输出。

1) GCN　对于第 l+1 层的文本图表示，计算为

$$H_{gcn}^{l+1}=\rho(\bar{A}H^{l}W^{l+1}) \tag{12}$$

其中，$H^l\in R^{m\times d}$ 为第 l 层的文本表示输出，$H^0=X$，W^{l+1} 为可学习的参数矩阵，ρ 为 Leaky_relu 激活函数。

2) ARS　首先对前 $l+1$ 层的所有文本表示输出与当前层的文本图表示分配注意力分数，计算为

$$H_{total}^{l+1}=[H^0,H^1,...,H^l,H_{gcn}^{l+1}] \tag{13}$$

$$H_{mean}^{l+1}=\mathrm{Mean}(H_{total}^{l+1}) \tag{14}$$

$$L^{l+1}=\sigma(W_l H_{mean}^{l+1}+b) \tag{15}$$

其中，$H_{total}^{l+1}\in R^{(l+2)\times m\times d}$、$H_{mean}^{l+1}\in R^{(l+2)\times d}$ 为不同维度的文本表示，$L^{l+1}=[\lambda^0,\lambda^1,...,\lambda^l,\lambda_{gcn}^{l+1}]$ 为各文本表示的注意力分数，W_l 与 b 为可学习的参数矩阵，σ 为 sigmoid 函数。

随后，使用注意力分数和残差结构得到当前层的文本表示输出 H^{l+1}，计算为

$$H^{l+1}=\lambda^0H^0+\lambda^1H^1+...+\lambda^lH^l+\lambda_{gcn}^{l+1}H_{gcn}^{l+1} \tag{16}$$

2.3.3 读出层

如图 1 所示，读出层利用注意力机制聚合单词特征，得到最终文本表示，并预测文本类别。最终文本表示 h_G 计算为

$$h_i=\sigma(h_i^{l+1}W_s+b_s)\odot\psi(h_i^{l+1}W_t+b_t) \tag{17}$$

$$h_G=\frac{1}{|V|}\sum_{i=0}^{m}h_i+\mathrm{Maxpooling}(h_1,..,h_m) \tag{18}$$

其中，σ 为 sigmoid 函数，$\sigma(\cdot)$ 表示对单词分配重要性系数，ψ 为 tanh 函数，$\psi(\cdot)$ 表示对单词特征进一步转换，W 与 b 为可学习的参数矩阵。除此之外，为了发挥每个词和重要词的作用，提取平均特征和重要特征，得到最终文本表示 h_G。

最后，使用 soft max 函数预测文本类别，目标函数为交叉熵损失函数，并使用正则项 RWACD，计算过程为

$$\bar{y}_G=\mathrm{softmax}(W_yh_G+b_y) \tag{19}$$

$$L=-\sum\nolimits_i y_G\log(\bar{y}_G)+\xi\times\mathrm{RWACD} \tag{20}$$

其中，$\bar{y}_G$ 为预测的文本类别，W_y 与 b_y 为可学习的参数矩阵，y_G 为真实的文本类别，ξ 为 RWACD 的系数。

3　实验部分

3.1　实验环境

文本算法的实验环境如表 1 所示。

表 1　实验环境

Tab. 1　Experimental environment

实验环境	环境配置	实验环境	环境配置
操作系统	Ubuntu 20.04.3	编程语言	Python 3.7.11
显卡	Nvidia GTX 2060S	开发工具	Pycharm
CUDA 版本	11.4	深度学习框架	Tensorflow 2.4.1

3.2　数据集

本文考虑使用以下 6 个数据集测试 RA-GCN 的性能，表 2 展示了数据集的统计数据，其中*表示该数据集未给出验证集。

表 2　数据集信息

Tab 2　Dataset information

数据集	训练集	验证集	测试集	类别	平均长度
MR*	7108	-	3554	2	18.46
Tnews	53360	10000	5000	15	12.01
Ohsumed*	3357	-	4043	23	79.49
R8*	5485	-	2189	8	41.25
SST-5	8544	1101	2210	5	17.75
SST-2	6919	871	1820	2	17.75

a) MR 数据集。含有正负面极性的 2 分类英文情感数据集。

b) Tnews[20]数据集。15 个类别的中文新闻分类数据集。

c) Ohsumed 数据集。23 个类别的英文心血管疾病医学摘

要分类数据集。

d) R8 数据集。8 个类别的路透社英文新闻分类数据集。

e) SST-2、SST-5 数据集。分别为 2 分类、5 分类英文情感分类数据集。

3.3 基线

由于文献[14]的方法适用于节点分类，与本文关注的文本图分类方向不符合，因此本文仅考虑与以下基线进行比较：

a) 传统的深度学习文本分类方法。包括 TextCNN[21]和 TextRNN[22]。

b) 基于单张文本-单词异构图的文本分类方法。包括 TextGCN[5]和 TextSGC[7]。

c) 基于每个文本图表示的图分类方法。包括 Huang[8]和 RA-GCN，不含 RWACD 和 ARS 的 P-GCN，P-SGC。

3.4 参数设置

对于未给验证集的数据集，将训练集随机分成 9:1 的比例用于实际训练和验证。对于初始单词特征，英文使用 200 维的预训练 GloVe[23]词向量，中文采用文献[24]中通过搜狗新闻训练的 300 维词向量。词汇外(Out of Vocabulary，OOV)单词从均匀分布[-0.01，0.01]中随机采样得到。算法使用 Adam[25]优化器，学习率设置为 0.001，其余参数根据不同数据集调整。模型性能使用准确度(Accuracy)进行衡量。

3.5 实验结果

表 3 为各模型在 6 个数据集上的准确度表现，实验结果为各模型训练 5 次的平均值。可以看出，RA-GCN 均取得了最好的结果。

表 3 实验结果

Tab. 3 Experimental results

模型	MR	Ohsumed	R8	SST-5	SST-2	Tnews
TextCNN	0.7775	0.5844	0.9571	0.423	0.8049	0.5602
TextRNN	0.7768	0.4927	0.9631	0.4263	0.8038	0.5518
TextGCN	0.7674	0.6836	0.9707	0.4063	0.8102	-
TextSGC	0.759	0.685	0.972	-	-	-
Huang	-	0.694	**0.978**	-	-	-
P-GCN	0.7853	0.684	0.9757	0.4376	0.8313	0.5684
P-SGC	0.783	0.6852	0.9689	0.4384	0.8264	0.5672
RA-GCN	**0.796**	**0.695**	**0.978**	**0.462**	**0.8451**	**0.5724**

对比传统方法，CNN 和 RNN 在大部分数据集上的性能均不如基于图的方法，证明了图模型有利于文本分类。对比基于每个文本图分类的模型 Huang、P-GCN、P-SGC、RA-GCN 和基于单张异构图分类的模型 TextGCN、TextSGC，前者在大多数情况下均优于后者，特别是在 MR、SST-2 等短文本数据集上，验证了基于每个文本图表示的图分类方法的有效性。

在 6 个数据集上的结果证明了所提文本分类算法 RA-GCN 的性能。RA-GCN 在 MR、SST-2、SST-5 和 Tnews 短文本数据集上提升较大，在长文本数据集上提升较小。因为构造的实际文本图拓扑并非真实潜在的文本拓扑结构，然而由于短文本的图规模较小，在 GCN 消息传递机制的作用下，单词信息传播广泛且迅速，RWACD 和 ARS 能够抑制过度平滑现象和弥补由于图拓扑差异引起的文本信息损失，所以 RA-GCN 能学习到更准确的文本表示。但是长文本的图规模较大，拓扑差异导致信息的传播速度不像小规模图一样流畅，造成了模型学习不到准确的文本表示，RWACD 及 ARS 发挥的作用较小，因此 RA-GCN 在长文本数据集上的文本分类性能提升不显著。

3.6 对比实验及过平滑现象分析

本小节以 GCN、SGC 为基础，验证 RWACD、ARS 对提升模型性能和抑制过平滑现象的作用，分析过平滑现象。实验均在 MR、SST-5 数据上进行，并抽取了 4 条 MR 测试集中的样本用于部分实验结果的可视化和分析，样本描述如表 4 所示。

表 4 样本描述

Tab. 4 Sample description

序号	样本(0/1 类)
1	a magnificent drama well worth tracking down(1 类)
2	an awkwardly contrived exercise in magic realism(0 类)
3	i'll put it this way if you're in the mood for a melodrama narrated by talking fish, this is the movie for you(1 类)
4	children and adults enamored of all things pokemon won't be disappointed(0 类)

构造分别含有 RWACD 或 ARS 的模型 RW-GCN、RW-SGC、ARS-GCN、ARS-SGC，不含及含有 RWACD 和 ARS 的模型 P-GCN、P-SGC、RA-GCN、RA-SGC。观察各模型在 MR 和 SST-5 数据集上的性能表现，探讨 RWACD 和 ARS 对模型性能的影响和在样本上的表现。最后分析了本文关注的文本图分类方向的过平滑现象。为了更好的区分各模型的表现，使用不同符号表示不同模型，模型说明如表 5 所示。

表 5 模型说明

Tab. 5 Model description

模型	简称	符号	含有结构	
			RWACD	ARS
P-GCN、P-SGC	P-模型	■	否	否
RW-GCN、RW-SGC	RW-模型	▲	是	否
ARS-GCN、ARS-SGC	ARS-模型	◆	否	是
RA-GCN、RA-SGC	RA-模型	▼	是	是

3.6.1 RWACD 与 ARS 的作用

1) RWACD、ARS 对模型分类性能的影响

表 6 为 8 个模型在 MR、SST-5 测试集上的文本分类准确度表现，实验结果为训练 3 次的平均值。

表 6 对比实验结果

Tab. 6 Comparative experimental results

模型	MR	SST-5
P-GCN	0.784	0.4376
P-SGC	0.7803	0.438
RW-GCN	0.7873	0.4429
RW-SGC	0.7847	0.4434
ARS-GCN	0.7943	0.4585
ARS-SGC	0.7926	0.4578
RA-GCN	0.7955	0.4623
RA-SGC	0.794	0.4586

从文本图构造方式的角度看，基于 GCN 和 SGC 提出的 8 个模型分类性能均优于 TextGCN 和 TextSGC 模型，这突出了基于文本图数据的文本图分类方法的优点。从是否含有 RWACD 和 ARS 的角度看，在 MR 和 SST-5 的实验结果中，不含 RWACD 和 ARS 的 P-模型分类性能最差，含有 RWACD 的 RW-模型较 P-模型有略微提升，证明了 RWACD 能够提升模型分类性能。含有 ARS 的 ARS-模型性能在 MR 和 SST-5 数据集上表现优异，取得了比 P-模型和 RW-模型更突出的分类性能，这凸显了从图拓扑角度优化模型性能决策的正确。含有 ARS 和 RWACD 的 RA-模型分类性能最优，在 MR 和 SST-5 上取得了最好的分类效果，这证明了 RWACD 和 ARS 能够同时提升模型的分类性能。

2) RWACD、ARS 对不同层数模型性能的影响

图 2 为各模型在 MR 测试集上的准确度和 WACD 随层数的变化曲线。可以看出，P-模型在分类性能和 WACD 上的表现均取得最差；在图 2 中，随着层数增加，P-模型的 WACD 逐渐下降，分类性能先增加后持续下降，说明一定程度的平滑可以提升模型性能，但是执行多次平滑后，会对模型性能

带来影响。对比 P-模型，RW-模型的分类性能和 WACD 略微提升，说明 RWACD 能够降低图数据的过度平滑，提升模型性能。ARS-模型的分类性能和 WACD 较 P-模型和 RW-模型提升明显，说明从图拓扑角度弥补文本信息损失能够显著提升模型性能和抑制过平滑现象。RA-模型的分类性能和 WACD 取得最佳，这说明 RWACD 和 ARS 能同时提升模型分类性能和抑制过平滑现象。

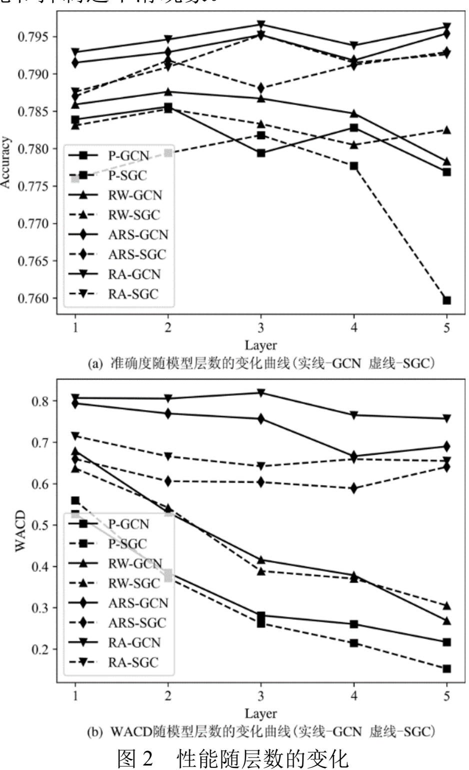


图 2　性能随层数的变化

Fig. 2　Accuracy and WACD changes with the number of layers

3) ARS 的作用分析

在上述两个实验中，ARS-模型性能表现突出，这是因为 ARS 从图拓扑角度出发，弥补了由于图拓扑差异带来的文本信息损失。为了更深一步证明 ARS 的作用，本小节设计了针对 ARS 的对比实验，探讨在破损的文本图数据上，模型是否能达到或接近在未破损图数据下模型的性能。

为了突出 ARS 的作用，以 2 层 P-GCN、ARS-GCN 为基线，去除读出层的注意力机制，随机删除文本图的边以破坏图拓扑，并逐渐提高删除比例。与文献[19]不同，对包括测试集的所有文本执行上述操作，并在训练过程中保持拓扑结构不变。为了突出实验结果，P-GCN 实验的删除比例最高为 20%，ARS-GCN 为 50%，两个模型在 MR 测试集上的实验结果如图 3 所示。

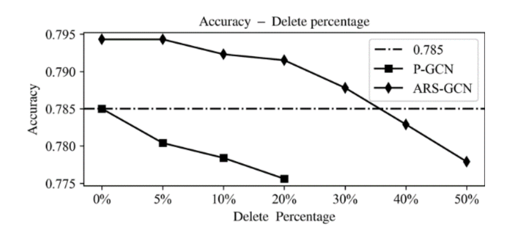


图 3　不同删除比例下的模型性能

Fig. 3　Model performance with different deletion ratios

从图中看出，在未破损的文本图数据上，ARS-GCN 模型性能明显优于 P-GCN，说明人为构造的文本图拓扑与真实潜在的文本图拓扑存在偏差，这验证了 2.2 节中 ARS 提出的初衷。随着删除比例的提高，不含 ARS 的 P-GCN 模型的性能急剧下降，然而对于 ARS-GCN 模型，尽管删除比例到达 30%左右，模型性能依旧能抵达或超越 P-GCN 模型的性能，说明了 ARS 能够弥补由于图拓扑差异带来的文本信息的损失，这再一次验证了从图拓扑角度出发优化模型性能决策的正确。

单从 ARS-GCN 曲线可以看出，删除比例在超过 20%后，模型性能急剧下降，这是因为高的删除比例会产生一些与其他节点无边连接的孤立节点，这种节点与其他节点无信息交互，造成了图模型捕捉不到词共现信息，学习不到准确的文本表示，因此造成了模型性能急剧下降；然而在 0%-20%的删除比例下，孤立节点的产生概率小，但是依旧对模型性能产生了影响，然而 ARS 弥补了由于图拓扑差异带来的文本信息损失，模型性能依旧可以到达或接近在原始数据下 ARS-GCN 模型的性能。

为了更清晰的观察 ARS 的表现，本小节探讨了两个模型在表 4 样本上的分类表现。ARS-GCN 的数据删除比例为 30%，P-GCN 不设置删除比例，分类结果如表 7 所示，其中√表示预测正确，×表示预测错误，结果为模型训练 3 次的平均值。

表 7　4 个样本的模型预测结果

Tab. 7　Model prediction results for 4 samples

样本	类别	0/1 类预测概率(是否预测正确)	
		P-GCN(原始数据)	ARS-GCN(删除比例 30%)
1	1	0.000/**1.000**(√)	0.022/**0.978**(√)
2	0	**0.996**/0.004(√)	**0.973**/0.027(√)
3	1	0.239/**0.761**(√)	0.384/**0.616**(√)
4	0	0.358/**0.642**(×)	**0.810**/0.190(√)

从表 7 中看出，ARS-GCN 对前 3 条样本的类别概率预测结果接近未设置删除比例的 P-GCN 结果；然而在第 4 条数据上，ARS-GCN 预测正确，P-GCN 预测错误，ARS-GCN 的预测结果要优于 P-GCN。这说明了在 ARS 的作用下，删除比例在 30%条件下的 ARS-GCN 的性能表现接近甚至超越了原始数据下 P-GCN 的性能，这从真实样本角度验证了图 3 中的实验结果。

综上所述，人为构造的文本图拓扑与潜在文本真实图拓扑之间存在差异，ARS 结构能够弥补由于这种差异带来的文本信息的损失，提升模型性能。

4) 案例分析

在表 4 样本的基础上，本小节在 2 层 P-GCN 和 RA-GCN 条件下，可视化了第 1、2 条样本内单词与其他单词的平均距离(图 4)；可视化了第 1、2 条样本 ACD 值随层数的变化曲线(图 5)；可视化了不同层数的 P-GCN 和 RA-GCN 对第 3、4 条样本的预测结果(表 8)，其中√表示模型预测正确，×表示预测错误。

表 8　第 3 和第 4 条样本的模型预测结果

Tab. 8　Model predictions for sample 3 and sample 4

模型(样本)	不同层数模型的预测结果				
	1	2	3	4	5
P-GCN(Sanple3)	√	√	×	×	√
ARS-GCN(Sanple3)	√	√	√	√	√
P-GCN(Sanple4)	√	×	√	√	×
ARS-GCN(Sanple4)	√	√	√	√	√

从图 4 中看出，RA-GCN 模型显著提升了单词间的平均距离，例如单词 worth 与其他单词的平均距离从 P-GCN 的 0.12 上升为 RA-GCN 的 0.61。在图 5 中 P-GCN 结果中，样本的 ACD 值在第三层接近于 0，单词之间变的相似，符合文献[12]中所描述的过平滑现象；然而在 RA-GCN 的结果中，样本的 ACD 值提升明显，说明 RWACD 和 ARS 抑制了过平滑现象。在表 8 样本 3、4 的结果中，P-GCN 预测正确 3 次，RA-GCN 全部预测正确，说明 RWACD 和 ARS 提升了模型对样本的分类性能。

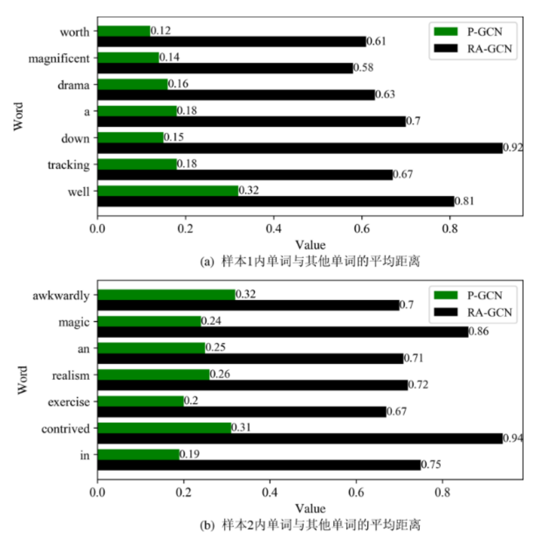


图 4　单词与其他单词的平均距离

Fig. 4　Average distance of words from other words words

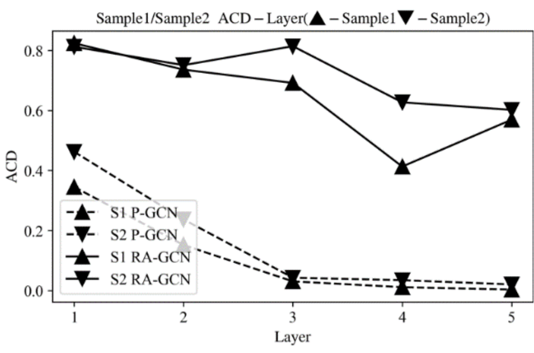


图 5　样本 ACD 值的变化曲线图

Fig. 5　The change curve of the sample ACD value

3.6.2 过平滑现象分析

从 3.6.1 节案例分析中的图 5 观察到，3 层的 P-GCN 就已经使样本的 ACD 趋近于 0，两个样本均出现了文献[12]描述的过平滑现象；并且在图 2 中观察到，随着层数的堆叠，P-模型的 WACD 逐渐下降，分类性能先上升后持续下降，说明一定程度的平滑有利于提升文本分类性能，但执行多次平滑后会损害模型分类性能。为此，本文假设：本文所关注的基于每个文本图表示的文本图分类领域，随着网络层数的堆叠，数据集内部分样本出现过平滑现象，且随着层数的堆叠，出现过平滑现象的文本越来越多，影响了模型的分类性能。

为了验证上述假设，借助构造的 8 个模型，分析在 MR 和 SST-5 测试集中文本图 ACD 小于某阈值时的文本数随网络层数的变化。实验结果如图 6 所示，为了突出部分模型的性能，模型之间的阈值取值不同，阈值取值已在图中标注，其中(a)(b)图为 GCN、SGC 在 MR 测试集上的结果，(c)(d)图为在 GCN、SGC 在 SST-5 测试集上的结果。

结合图 2 和图 6 中 P-模型的实验结果看出，ACD 小于 0.3 的文本数随网络层数的上升逐渐增多，WACD 随层数的上升逐渐下降，分类性能先上升后下降，结合图 4、图 5 中 P-GCN 的可视化结果，说明：本文所关注的基于每个文本图表示的图分类领域，过平滑现象体现在以文本图为单位的文本表示中，这种过平滑现象在浅层网络就已出现，并且随着网络层数的堆叠，出现过平滑现象的文本逐渐增多，过平滑现象愈加明显，但是模型性能在 2-3 层时才开始出现下降。

从 RW-模型、ARS-模型的曲线看出，RWACD 和 ARS 均能减少出现过平滑的文本数，抑制过平滑现象，提升模型性能。

RA-模型的结果均取得最佳，说明 RWACD 和 ARS 同时减少了出现过平滑的样本数，抑制了过度平滑现象，提升了模型分类性能。

综上所述，以 P-模型为基准，本文关注方向的过平滑现象以文本图为单位，在浅层网络就已出现，且过平滑文本数随着网络堆叠而逐渐增加，损害了模型性能；RWACD 和 ARS 均能减少过平滑样本数，抑制过平滑现象，提升模型分类性能。

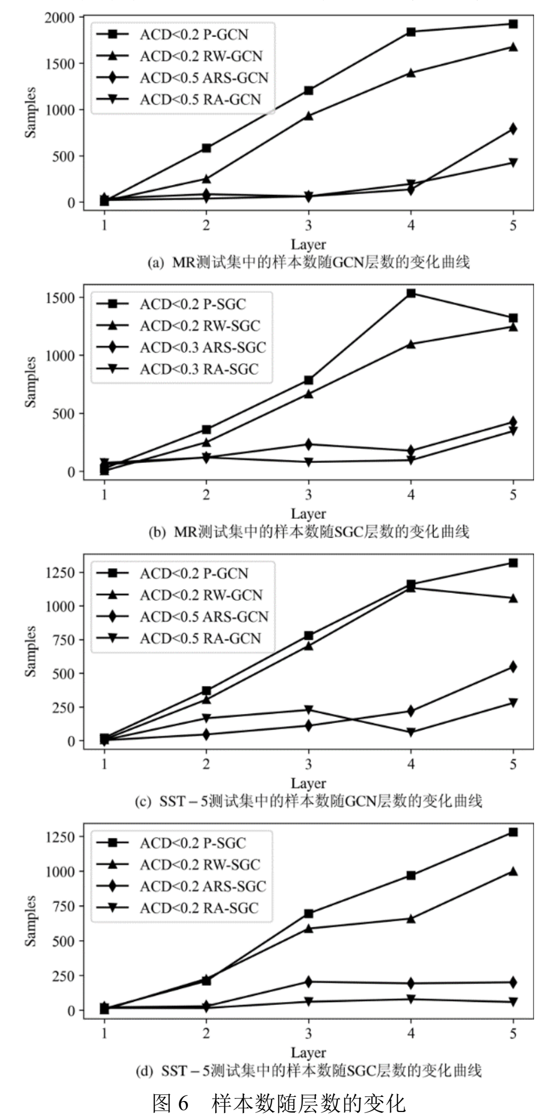


图 6　样本数随层数的变化

Fig. 6　Variation of the number of samples with the number of layers

4　结束语

本文提出了适用于多个文本图表示的平滑度衡量指标加权平均余弦距离 WACD，提出了抑制过度平滑的正则项 RWACD。提出了注意力和残差的网络结构 ARS，弥补由于文本图拓扑差异引起的图表示学习带来的文本信息的损失，同时抑制过度平滑现象。提出了基于 RWACD 和 ARS 的图卷积神经网络文本分类算法 RA-GCN。在 6 个数据集上证明了 RA-GCN 的性能，并且通过多个对比实验验证了 RWACD 和 ARS 的作用。

参考文献：

[1] Li Qian, Peng Hao, Li Jianxin, *et al.* A survey on text classification: from shallow to deep learning [J/OL]. ACM Trans on Interactive Intelligent Systems, 2021, 37 (4) . (2021-04) [2021-12-11]. https://arxiv.org/pdf/2008. 00364. pdf.

[2] Kowsari K, Jafari M K, Heidarysafa M, *et al.* Text classification algorithms: a survey [J]. Information, 2019, 10 (4): 150.

[3] Chiu B, Sahu S K, Sengupta N, *et al.* Attending to inter-sentential features in neural text classification [C]// Proc of the 43rd International ACM SIGIR Conference on Research and Development in Information

Retrieval. New York: ACM, 2020: 1685-1688.

[4] 何力，郑灶贤，项凤涛，等．基于深度学习的文本分类技术研究进展 [J]．计算机工程，2021, 47 (2): 1-11. (He Li, Zheng Zaoxian, Xiang Fengtao, *et al.* Research progress of text classification technology based on deep learning [J]. Computer Engineering, 2021, 47 (2): 1-11.)

[5] Yao Liang, Mao Chengsheng, Luo Yuan. Graph convolutional networks for text classification [C]// The 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33 (1): 7370-7377.

[6] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [C/OL]/ The 5th International Conference on Learning Representations. 2017. (2017-02) [2021-12-11]. https://arxiv. org/pdf/1609. 02907. pdf.

[7] Wu, F, Zhang Tianyi, Souza A, *et al.* Simplifying graph convolutional networks [C]// Proc of the 36th International Conference on Machine Learning. [S. I.] : PMLR, 2019: 6861-6871.

[8] Huang Lianzhe, Ma Dehong, Li Sujian, *et al.* Text level graph neural network for text classification [C]// Proc of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2019: 3442-3448.

[9] Zhang Yufeng, Yu Xueli, Cui Zeyu, *et al.* Every document owns its structure: inductive text classification via graph neural networks [C]// Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 334-339.

[10] Li Yujia, Tarlow D, Brockschmidt M, *et al.* Gated graph sequence neural networks [C/OL]/ The 4th International Conference on Learning Representations. 2016. (2016) [2021-12-11]. https://arxiv. org/pdf/1511. 05493. pdf.

[11] 范国凤，刘璟，姚绍文，等．基于语义依存分析的图网络文本分类模型 [J]．计算机应用研究, 2020, 37 (12): 3594-3598. (Fan Guofeng, Liu Gui, Yao Shaowen, *et al.* Text clasification model with graph network based on semantic dependency parsing [J]. Application Research of Computers, 2020, 37 (12): 3594-3598.)

[12] Li Qimai, Han Zhichao, Wu Xiaoming. Deeper insights into graph convolutional networks for semi-supervised learning [C]// Proc of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 3538-3545.

[13] Cai Chen, Wang Yusu. A note on over-smoothing for graph neural networks [EB/OL]. (2020-06-23) [2021-12-11]. https://arxiv. org/pdf/2006. 13318. pdf.

[14] Chen Deli, Lin Yankai, Li Wei, *et al.* Measuring and relieving the over-smoothing problem for graph neural networks from the topological view [C]// The 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 3438-3445.

[15] Wang Guangtao, Ying R, Huang Jing, *et al.* Multi-hop attention graph neural networks [C]// Proc of the 30th International Joint Conference on Artificial Intelligence. [S. I.] : ijcai. org, 2021: 3089-3096.

[16] Yang Tianmeng, Wang Yujing, Yue Zhihan, *et al.* Graph pointer neural networks [EB/OL]. (2021) [2022-01-05]. https://arxiv. org/pdf/2110. 00973. pdf.

[17] Vinyals O, Fortunato M, Jaitly N. Pointer networks [C]// Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems. 2015: 2692-2700.

[18] Li Guohao, Muller M, Thabet A, *et al.* DeepGCNs: Can GCNs go as deep as CNNs? [C]// Proc of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 9266-9275.

[19] Rong Yu, Huang Wenbing, Xu Tingyang, *et al.* DropEdge: towards deep graph convolutional networks on node classification [C/OL]/ The 8th International Conference on Learning Representations. 2020. (2020-05-12) [2022-01-05]. https://arxiv. org/pdf/1907. 10903. pdf.

[20] Xu Liang, Hu Hai, Zhang Xuanwei, *et al.* CLUE: A Chinese language understanding evaluation benchmark [C]// Proc of the 28th International Conference on Computational Linguistics. [S. I.] : International Committee on Computational Linguistics, 2020: 4762-4772.

[21] Kim Y. Convolutional neural networks for sentence classification [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1746-1751.

[22] Liu Pengfei, Qiu Xipeng, Huang Xuanjing. Recurrent neural network for text classification with multi-task learning [C]// Proc of the 25th International Joint Conference on Artificial Intelligence. [S. I.] : IJCAI/AAAI Press, 2016: 2873-2879.

[23] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1532-1543.

[24] Li Shen, Zhao Zhe, Hu Renfen, *et al.* Analogical reasoning on Chinese morphological and semantic relations [C]// Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 138-143.

[25] Kingma D P, Ba J L. Adam: A method for stochastic optimization [C/OL]/ The 3rd International Conference on Learning Representations. 2015. (2015) [2022-01-05]. https://arxiv. org/pdf/1412. 6980. pdf.